Cluster Analysis of Stocks

Team Members: Jordan Zane, Caleb Clough, Matt Dyer, Evan Linden

Description

This project will aim to mine extensive longitudinal American stock market data to identify interesting trends.

In particular, our group aimed to cluster individual stocks into groups to see what individual stocks tend to “move” together.

This analysis was set to uncover greater nuance and cross relationships outside of typical stock groupings (technology, manufacturing, finance, etc.) and allow for novel (or atleast modified) classification grouping schemes.

Our Dataset

Our main dataset is a .csv containing 5 years of AMEX, NYSE, and NASDAQ end of day data. The source of this data is Kaggle and the URL is as follows:

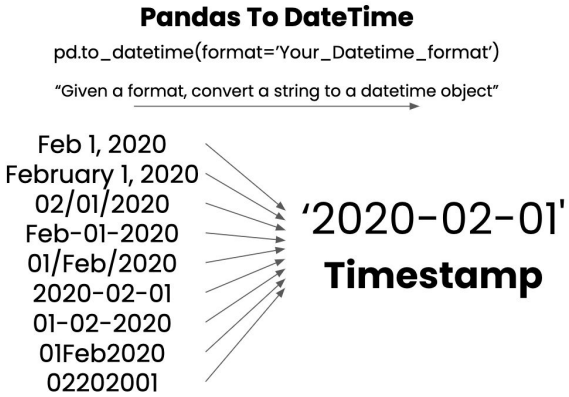
https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories?select=fh_5yrs.csv

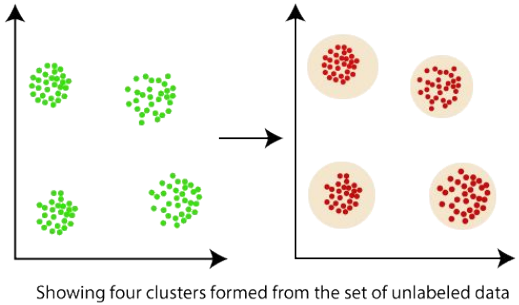
The set of data has been frozen, downloaded, and stored on the cloud at the following location:

https://drive.google.com/drive/folders/17RyZ9pbuVMTZXn2SvbAhO8fvzA-08Kbv?usp=sharing

Data Preparation Work

- Data Cleaning
  - Drop any NULL data values, if there are any
  - Check and replace any string values from the data set (e.g. $), and ensure data usable is in date time.
  - Examine for any outliers and remove if necessary
- Data Preprocessing
  - Add an attribute to identify whether a certain stock increased day over day.
  - Reduce the size of the dataset using a data reduction technique
    - Use date-aligned data to implement clustering of stocks
    - Implement data cube



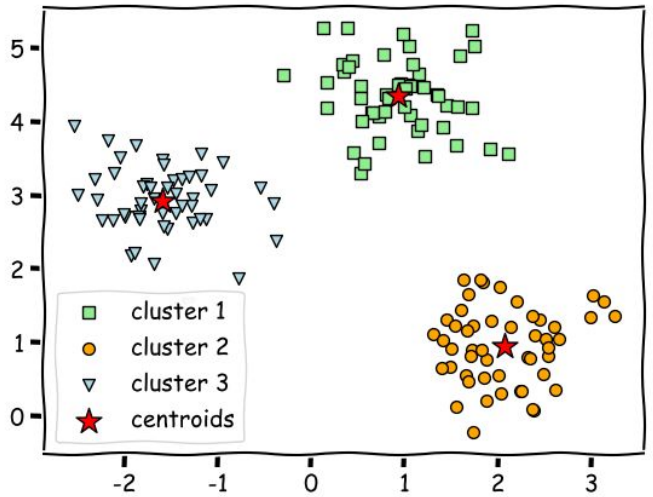

Showing four clusters formed from the set of unlabeled data

List of Tools

- Python
- NumPy
- Matplotlib
- Pandas
- Sklearn
- GitHub
    - https://github.com/evanlinden/Data-Mining-Group-02.git

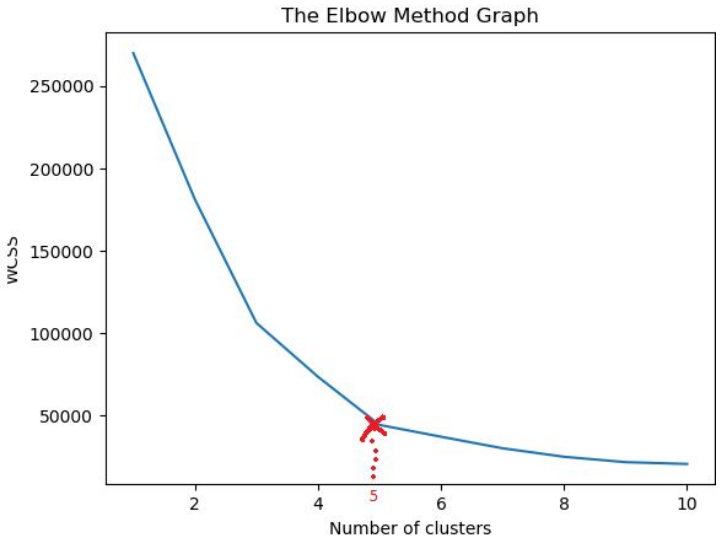
Classification/Clustering/ETC Applied

- Data Integration
  - Kaggle dataset updated to 6/14/20
  - Merge auxiliary datasets (if necessary)
  - Use stock symbols (eg. TSLA) for entity identification
- Data Analysis
  - Use daily stock movement (close - open)
  - Clustering to identify similar stocks
    - Identify number of clusters
    - K-means clustering



Evaluation

- Stock metadata intuition (eg Are tech companies in same clusters?)
- Assess similarity in stock movement among clusters and between clusters
  - Move in same direction?
  - Move on same days?
- Distortion score in k-means clustering
- Elbow method for optimal number of clusters



Knowledge Gained

The results from the clustering show clusters of stocks that are strongly related to each other based on their classification. Our clustering showed that stocks aren't necessarily related to each other by their specific sector, but more by their genre. For example, we found that GM and Ford aren't clustered together with any other car companies, but rather they are found in a cluster with steel production and mining companies. This is easily explained by the fact that cars use a lot of steel and car prices move with spikes of intermediate goods such as steel. There is a lot to be said about our results and how much information there is to be gained about the stock market from them. As a group we are really happy to see that the clusters do in fact make sense and are interesting enough to be applied to real life.

Applying That Knowledge

Obviously the idea around analyzing the stock market is to be able to make predictions about stocks and how they will move. The goal of clustering for us was to separate the different tickers out into their clusters so that we could buy stocks based on their certain clusters price movement. This also gives us the ability to better diversify our portfolios so that we can maximize our financial gains. Using the knowledge we have from this clustering we can test it out and see how it works in the long run. There are so many stocks to be discovered with this knowledge and we are sure it will be fun!